Transcending the Dilemma of Applying IHL to LAWS in Agency Spillover Engagements: A Legal Analytical Perspective on Human-Machine Hybrid Agency

Fan Yang[*]

ABSTRACT

The militarization of Lethal Autonomous Weapons Systems inevitably triggers "agency spillover engagements", situations where algorithmic agency deviates from human presets. This has led to systemic applicability ruptures across International Humanitarian Law, state responsibility, and International Criminal Law. This paper argues that the root cause of the failure lies in the strict adherence to the human moral agent presumption. To address this, this paper proposes an ontological perspective of human-machine hybrid agency and employs the principal-agent relationship as an analytical lens. It further derives three domain-specific tools – the Joint Cognitive System, Risk Allocation, and Reckless Delegation – to transcend these dilemmas within lex lata *through evolutionary and teleological interpretation.*

Keywords: Lethal Autonomous Weapons Systems; Agency Spillover Engagement; Hybrid Agency; Principal-Agent; International Humanitarian Law; Legal Interpretation

I. Introduction

Compared to all previous weapon systems, Lethal Autonomous Weapons Systems (LAWS)[1] exhibit one fundamental difference: due to the intervention of Artificial Intelligence (AI), LAWS intrinsically possess a certain degree of autonomous agency.[2] Similar to civil or commercial AI systems that have rapidly evolved over

---

[*] Associate Professor of International Law, Deputy Director of Cyberspace International Law Center, Xiamen University School of Law, China.

[1] The scope of military AI is analytically broader than LAWS, including using AI capability in military cyber operations, as well as AI-enabled decision support system in military operations. See Fan Yang, *Applying International Law to AI in the Military Domain: An Integrated Approach*, GC REAIM Expert Policy Note Series, April 2025, available at: https://hcss.nl/wp-content/uploads/2025/04/Yang.pdf. (all internet references were accessed on May 6, 2026). This paper's focus is on LAWS.

This paper retains the term "LAWS" as it remains the established designation in the CCW GGE process. The author acknowledges ongoing debate about whether the qualifier "lethal" is clearly defined or analytically necessary; the framework proposed here applies to autonomous weapon systems with lethal consequences regardless of the terminological convention adopted.

[2] The distinction between "autonomous" and "automatic" (or "automated") is foundational to defining LAWS. A merely automatic system follows predetermined, deterministic rules; an autonomous system can modify its behaviour in response to environmental conditions. See Paul Scharre, "*Army of None: Autonomous Weapons and the Future of War*", W.W. Norton, New York, 2018, pp. 27-45; see also United States Department of Defense, Directive 3000.09, "Autonomy in Weapon Systems", 25 January 2023, defining autonomous weapon systems as those that "can select and engage targets without further

the past two years (e.g., agentic AI, autonomous driving, embodied intelligent robots),[3] these AI weapon systems possess an "entity capacity" to autonomously plan, interact, and alter the state of reality in digital or physical environments with minimal human intervention.[4] The law needs to properly incorporate this agency into its analytical framework, just as it has historically succeeded in addressing the endogenous moral agency of humans, the biological agency of animals,[5] and the fictive group agency[6] of organizations such as corporations and partnerships. Ignoring it or failing to assimilate it will inevitably lead to legal dilemmas. This is because AI agency can disrupt or even sever the legal analysis of human responsibility that would apply in equivalent or similar situations without AI. Human subjects will frequently argue as a defence that, under the current legal framework, liability cannot be attributed to them because the harmful consequences were caused by AI agency.

An ideal research conclusion concerning LAWS and International Humanitarian Law (IHL) should be able to effectively address this agency,[7] analysing the determination of illegality and the pursuit of legal responsibility resulting from the use of LAWS during engagements – namely, whether there is a violation of IHL rules on the conduct of hostilities, and whether it triggers state responsibility and/or individual international criminal responsibility. Regarding the issue of legal responsibility, the scope of legal analysis transcends the specific branch of IHL, entering into the realms of the Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA) or International Criminal Law (ICL). However, these inquiries remain inextricably linked to, or premised upon, the IHL analysis.

---

intervention by a human operator." For a comparative analysis of definitions, see Filippo Santoni de Sio and Jeroen van den Hoven, "Meaningful Human Control over Autonomous Systems: A Philosophical Account", *Frontiers in Robotics and AI*, Vol. 5, No. 15, 2018.

[3] Many general discussions on the legal liability of AI also choose to approach the issue from the perspective of agency. See, e.g., Cullen O'Keefe, Ketan Ramakrishnan, Janna Tay and Christoph Winter, "Law-Following AI: Designing AI Agents to Obey Human Laws", *Fordham Law Review*, Vol. 94, 2025, pp. 57-129.

[4] Luciano Floridi defines agency through three dimensions: the capacity to interact with and affect the environment (interactivity), the capacity to change state independently of external control (autonomy), and the capacity to change its own rules of behaviour based on input (adaptability). See Luciano Floridi, "AI as Agency without Intelligence: On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis", *Philosophy & Technology*, Vol. 38, 2025, p. 30.

[5] See L. Floridi, above note 4, p.30.

[6] See Christian List, "Group Agency and Artificial Intelligence", *Philosophy & Technology*, Vol. 34, 2021, pp. 1213-1242.

[7] Military AI decision-support systems also present agency issues, but these primarily concern how human agency is reshaped or even eroded following the introduction of such new technologies. For research specifically discussing how this impact reflects on the legal analysis of IHL applicability. See Taylor Kate Woodcock, "Human/Machine(-Learning) Interactions, Human Agency and the International Humanitarian Law Proportionality Standard", *Global Society*, Vol. 38, No. 1, 2024, pp. 100-121.

Logically speaking, the legal crises that genuinely challenge the application of IHL in a dogmatic sense – which may be a factual reality or a defence invoked by potential responsible parties for exculpation – can be narrowed down to a specific category of scenarios. Due to the underlying agency of the weapon's AI system, compounded by complex interferences in real battlefield environments, LAWS may act beyond the preset instructions and rules of engagement (ROE) of human commanders or operators, resulting in consequences that allegedly violate IHL. This paper terms such scenarios Agency Spillover Engagements (ASE).[8] Conversely, when LAWS appear to follow the preset instructions and ROE of human commanders or operators, their inherent agency overlaps with, or can be deemed subsumed by, the agency of the relevant human subjects. Consequently, legal analysis needs only to address the latter, rendering the entire legal reasoning identical to traditional analytical pathways.

Examining the agency of LAWS and human subjects along a continuous spectrum helps explain numerous phenomena. For instance, the emphasis on ensuring "meaningful human control" in the context of LAWS,[9] or the demand for enhanced explainability during the research and development and design phases of weapon AI systems,[10] and other propositions like these all point toward a unified endeavour: the desire to institutionally limit and narrow the exercise of LAWS' agency, while expanding the dominance and suppression by human agency. Similarly, the consensus advocating for the prohibition of fully autonomous weapons[11] represents the international community's shared resistance against the extreme scenario where LAWS' agency completely marginalizes human agency. Despite these efforts, it must be acknowledged that as long as LAWS are introduced to the battlefield, the possibility of an agency spillover engagement inevitably exists. This anxiety does not stem from anthropomorphic imaginations found in sci-fi narratives of machines awakening and breaking free from human control; rather, it

[8] Some scholars refer to this as "attacks beyond preset parameters (in Chinese 超设定攻击)."See LENG Xinyu, "Principle of Meaningful Human Control under Topic of Lethal Autonomous Weapon System", *Journal of International Law (China)* Vol. 2, 2022, p. 17.

[9] For example, the principle of meaningful human control is expressed as making the utmost effort to implement human control over weapon systems during policy formulation, development, deployment, testing, use, and post-use assessment. The 2018 GGE report labelled these processes with numbers 0 to 5 and depicted them in a chart to facilitate understanding and policy adjustment by states. See Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/GGE.1/2018/3, Annex III, para. 10.

[10] See, e.g., International Committee of the Red Cross. (2021). *ICRC position and background paper on autonomous weapons systems*. International Committee of the Red Cross: Technical report, p. 7.

[11] This consensus is reflected in the "rolling text" of the Group of Governmental Experts (GGE) on LAWS under the Convention on Certain Conventional Weapons (CCW), as well as in the policy positions of major military AI powers and representative military AI enterprises. See, e.g., the Report of the 2023/2024 Session of the GGE on LAWS, and the *Blueprint for Action* endorsed at the 2024 REAIM (Responsible AI in the Military Domain) Summit.

is founded upon a stark technological reality.[12] When a system based on complex algorithms is granted a certain degree of autonomous decision-making authority in an open, chaotic battlefield environment, it will – when faced with unpredictable environmental stimuli – inevitably have a probability of generating emergent actions based on its underlying computational logic that a human commander could not predict *a priori*.

Consider the following hypothetical scenario: an autonomous loitering munition deployed for Suppression of Enemy Air Defences (SEAD) encounters an unknown electronic deception tactic. Instead of conventionally choosing to self-destruct or return to base, its algorithm autonomously learns and determines that a nearby civilian communication tower is a disguised interference source, subsequently initiating an attack to destroy it. The loitering munition does not experience a mechanical malfunction in the traditional sense; rather, while exercising its delegated authority, it executes an agency spillover engagement unforeseen by humans.

At this juncture, the core substantive obligations of IHL face the disintegration of their applicative foundation due to the absence of human empathy and subjective moral balancing. Simultaneously, because the attack deviates from specific State authorization and the direct intent (*mens rea*) of the commander, the attribution chain of State responsibility and the individual accountability mechanisms of International Criminal Law are dually obstructed.[13]

Since agency and specifically, its spillover, are the culprit creating this legal analytical dilemma, agency might also hold the key to resolving it. This paper attempts to use the agency of LAWS as a continuous thread of inquiry, drawing on both legal philosophy and philosophy of technology. The logical progression of this paper is as follows: First, it systematically examines the specific legal obstacles facing

---

[12] This is a technological inevitability. On the technical mechanisms through which machine learning systems produce unintended and harmful behaviour, including distributional shift (encountering environmental data outside the training distribution), reward hacking, and negative side effects, see Dario Amodei *et al.*, "Concrete Problems in AI Safety", arXiv preprint 1606.06565, 2016. For analysis of how these risks manifest specifically in lethal autonomous weapons systems, see Ingvild Bode and Tom Watts, "Technical Risks of (Lethal) Autonomous Weapons Systems", arXiv preprint 2502.10174, 2025, arguing that even rigorously tested systems may behave unpredictably and harmfully in real-world conditions due to the black-box nature of AI decision-making, susceptibility to reward hacking, goal misgeneralisation, and emergent behaviours that escape human control. See also Dustin Trusilo and David Danks, "Autonomous AI Systems in Conflict: Emergent Behavior and Its Impact on Predictability and Reliability", *Journal of International Humanitarian Legal Studies*, Vol. 14, No. 2, 2023, discussing how autonomous algorithmic systems operating in open environments produce nondeterministic, innovative and nontransparent behaviour at the level of data processing.

[13] See Andreas Matthias, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata", *Ethics and Information Technology*, Vol. 6, 2004, pp.175-183. The academic community has further subdivided this gap into the culpability gap, moral accountability gap, and active responsibility gap. See Filippo Santoni de Sio and Giulio Mecacci, "Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them", *Philosophy & Technology*, Vol. 34, 2021, pp.1057-1084.

IHL substantive rules and accountability mechanisms in agency spillover engagements, revealing the limitations of current interpretative pathways. Second, it traces and reconstructs the concept of agency in legal history and jurisprudence, arguing that during engagements utilizing LAWS, the functional agency of the AI-driven weapon system combines with the agency of human subjects to constitute a hybrid agency, which is precisely the core object that legal systems such as IHL must properly incorporate into their analysis. Finally, by developing conceptual tools related to agency, the paper explores how to transcend existing legal analytical barriers.

II. The Triple Legal Dilemma in Agency Spillover Engagements

In the ASE scenario, strictly adhering to the analytical pathways of legal dogmatics inevitably drives the substantive rules of IHL on the conduct of hostilities, the attribution logic of ARSIWA, and the individual accountability mechanisms under ICL into an impasse. Such agency spillover can be subsumed neither by the traditional legal concept of weapon malfunction nor by the doctrine of mistake of fact.

A. The Applicability Dilemma of Core Substantive IHL Rules

Under the dogmatic framework, the bearers of substantive IHL obligations are consistently human commanders or operators, a consensus that the international society has upheld till now.[14] The true legal crisis posed by LAWS in ASE lies in the fact that, the factual agency they exhibit substantively severs the traditional evaluative chain for determining whether a human subject has violated IHL obligations. Many IHL rules will have to answer this challenge, but here we focus on the three IHL principles on the conduct of hostilities.

The principle of distinction requires conflicting parties to distinguish between civilian and military objectives at all times.[15] In the traditional use of weapons, there exists a linear correspondence between a commander's intent to attack and the weapon's physical point of impact. However, deploying LAWS means outsourcing the cognitive process of target feature recognition and classification at the tactical edge to algorithms. When LAWS, operating in a complex electromagnetic environment, autonomously generate a decision to classify a civilian facility as a military objective and subsequently destroy it, the human's lawful initial intent and the unlawful final attack result are decoupled by the machine's agency.

[14] See Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/GGE.1./2019/3, Annex 4, p.14.
[15] Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 1125 UNTS 3, 8 June 1977 (entered into force 7 December 1978), Art. 48. (AP 1). "… the Parties to the conflict shall at all times distinguish between the civilian population and combatants and between civilian objects and military objectives and accordingly shall direct their operations only against military objectives."

Consequently, current IHL is unable to subsume this scenario within its traditional framework to hold the human commander in violation of the principle of distinction.

The principle of proportionality similarly faces the disintegration of its spatiotemporal foundation. This principle heavily relies on the decision-maker's subjective balancing of the "anticipated military advantage" against the "incidental damage" at a specific moment.[16] A commander's general estimation upon initiating deployment in the rear is vastly disconnected from the specific attack trajectory instantaneously generated by LAWS deep within the battlefield. Confronted with a spillover attack scenario temporarily created by machine agency, the commander arguably never conducted any "balancing" in the IHL sense, causing the evaluative chain to rupture at this exact juncture.

The principle of precautions requires the attacker to "cancel or suspend" an attack if it becomes apparent that the objective is not a military one.[17] This obligation is premised on the human possessing the physical and cognitive feasibility to comply. However, this faces an impenetrable physical barrier at the tactical end, as machine algorithms complete the OODA (Observe-Orient-Decide-Act) loop in milliseconds. When an ASE occurs, the commander is cognitively unable to penetrate the algorithmic black box and is physically excluded from the decision-making loop. Continuing to compel the commander to "suspend the attack in a timely manner" would contradict the fundamental legal maxim of *ultra posse nemo obligatur*, i.e. no one is obligated beyond what they are able to do.

B. The Applicability Dilemma of Attribution Rules in State Responsibility

The violation of IHL substantive obligations may lead to State responsibility, provided that such a violation can be attributed to a specific State under international law.[18] Arguably, the attributability of an IHL violation caused by the use of LAWS is self-evident. According to Article 4 of ARSIWA, an attack conducted by the armed forces of a State using any weapon system is naturally and directly attributable to the State, regardless of whether it uses a rifle, a missile, or an AI weapon. Therefore, there is no need to invoke other rules of attribution; the legal attribution judgement in the sense of State responsibility is clear and unquestionable.

---

[16] AP I, Art. 51(5)(b). Indiscriminate attacks are prohibited, including "an attack which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated."
[17] API, Art. 57(2)(b). "An attack shall be cancelled or suspended if it becomes apparent that the objective is not a military one or is subject to special protection or that the attack may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated".
[18] ARSIWA, Art. 2.

In the ASE scenario, a State might construct the following defence.[19] The attack by LAWS did indeed occur; however, the act is a manifestation of the weapon system's own agency, which constitutes an "independent intervening factor" (*novus actus interveniens*). Its connection to the operator is limited to the fact that the operator activated the weapon system, a fact that is a condition for the IHL violation rather than the proximate cause.[20] Since it cannot be regarded as the result of the operator's conduct, how can it then be attributed to the State? There are three alternative analytical approaches, each of which encounters interpretive difficulties that illuminate the broader challenge LAWS pose to the existing framework.

The first path attempts to attribute the autonomous conduct of LAWS directly to the State. Under Article 5 of ARSIWA, the conduct of a "person or entity" exercising elements of governmental authority shall be considered an act of the State. According to the International Law Commission's commentary, authorized entities can encompass a wide range of forms, such as public corporations, semi-public entities, and private companies, under the crucial condition that they are empowered by the internal law of the State.[21] Although LAWS objectively exhibit functional agency and can be considered to have been empowered to a certain extent, thus enabling them to autonomously search, identify, and engage targets; the critical issue is that, at least at the current stage, LAWS are not recognised as possessing legal personhood, and they cannot be directly interpreted as the "entity" referred to in Article 5.[22]

Then, could one invoke Article 8 of ARSIWA to treat the conduct of LAWS as conduct "directed or controlled" by a State? The application of Article 8 strictly requires the State to exercise "control" over the specific conduct. But in ASE, LAWS autonomously generate an attack trajectory beyond preset parameters based on underlying algorithms in a complex environment. At this point, the factual agency of the machine has transcended the State's initial boundaries of authorization. The State could argue in defence that the transgressive attack is neither the direct will of the State apparatus nor within the State's instantaneous control, but purely the result of the independently emergent agency of the algorithmic agent. Not to mention, the

---

[19] Some scholarly view holds that such a defence would not succeed for State-deployed autonomous capabilities. See Jonathan Kwik, "Autonomous Cyber Capabilities under International Law", CCDCOE (2021), p. 128; see also Heather A. Harrison Dinniss, "AI and Military Cyber Operations", in *Research Handbook on Warfare and Artificial Intelligence*, Edward Elgar Publishing, 2024, p. 274. This paper acknowledges and agrees with that conclusion. The defence is nevertheless constructed here in its strongest form, because doing so reveals precisely *where* the existing doctrinal framework struggles and *why* the hybrid agency perspective is needed to provide the theoretical grounding for the rejection.

[20] James Crawford, *State Responsibility*, Cambridge University Press, Cambridge, 2013, pp. 492-494.

[21] ILC Commentaries to the draft articles on Responsibility of States for internationally wrongful acts, extract from the Report of the International Law Commission on the work of its Fifty-third session, Official Records of the General Assembly, Fifty-sixth session, Supplement No. 10 (A/56/10), chp.IV.E.2, p. 92.

[22] For discussion that opposes anthropomorphizing LAWS, see, e.g., Bill Boothby, "AI Warfare and the Law", *International Law Studies*, Vol. 104, 2025, pp. 9-12.

interpretative standards of "effective control" [23] or "overall control" [24] have historically been used to address the relationship between States and armed opposition groups that are human organizations. Forcibly creating a legal fiction of AI as a highly organized non-State entity to mechanically apply Article 8 is a category mistake.

Another path attempts to build a bridge through the human operator. Since the operator belongs to the State's armed forces, falling under the definition of a State organ in Article 4 of ARSIWA, the State should bear responsibility as long as it is determined that it is the "conduct of" the operator that LAWS carries out an unlawful attack. However, in ASE, this transmission chain is similarly severed. When LAWS autonomously lock onto a target with millisecond decision-making speed and opaque black-box algorithms, the high-frequency agency of the machine substantively deprives the operator of cognitive and interventional capabilities at the tactical end. The operator can neither foresee the specific unlawful attack nor easily implement an effective suspension physically. As such, the unlawful engagement behaviour cannot be simply and directly evaluated in law as the conduct of the operator. The difficulty lies in determining the standard by which an autonomous weapon's attack beyond preset parameters is still considered that of the operator. Furthermore, whether the operator is responsible as a principal offender for the autonomous weapon's unlawful attack depends on an analysis of their subjective state. In other words, this analytical path entangles the question of State attribution with the determination of individual criminal responsibility. This pushes the legal analysis into the next thorny dilemma.

C. The Applicability Dilemma of Individual Accountability in ICL

Assuming that the autonomous weapon system itself does not involve a weapon whose use is prohibited or restricted by international law,[25] the problems arising in ASE manifest in two aspects when ICL attempts to pursue the individual criminal responsibility of relevant human subjects. The first is whether a causal link can be established between the result of the weapon system's attack violating IHL and the operator's conduct. The second is how to evaluate the psychological state of the operator. This corresponds to ICL's fundamental method of combining objective and subjective elements to attribute guilt.

---

[23] International Court of Justice, *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)*, Judgment (Merits), ICJ Reports 1986, para. 115; International Court of Justice, *Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v. Serbia and Montenegro)*, Judgment, ICJ Reports 2007, para. 400.

[24] International Criminal Tribunal for the former Yugoslavia, *Prosecutor v. Duško Tadić*, Case No. IT-94-1-A, Judgment (Appeals Chamber), 15 July 1999, paras. 116–141.

[25] When an autonomous weapon system involves a weapon whose use is prohibited by international law, the mere use of the autonomous weapon system already constitutes a war crime. These war crimes are conduct crimes, where a conviction can be secured without the need to consider the criminal outcome or the causal relationship that caused the criminal outcome.

Particularly concerning the subjective elements, LAWS create a degree of a responsibility vacuum. In ASE, although the commander has pressed the button to activate the system, which constitutes an objective act of deployment, they may not possess "knowledge" in the criminal law sense regarding the subsequent transgressive attack autonomously generated by the machine agent, let alone the "intent" to pursue such an unlawful consequence. Here, a rupture exists between the human's criminal intent (*mens rea*) and the machine's actual harmful active conduct (*actus reus*). It is difficult to constitute direct intent, amounting at most to indirect intent.

Per ICL jurisprudence, the subjective elements of war crimes generally encompass direct intent, with indirect intent existing only as an exception. International criminal justice has no precedent for holding a principal perpetrator criminally liable based on negligence, and only in limited circumstances does it recognise a situation where the defendant in a war crime case possessed the mental state of indirect intent.[26] Article 30 of the Rome Statute sets extremely stringent requirements for the subjective elements of war crimes, requiring the perpetrator to act with "intent and knowledge". It requires that the person "means to engage in the conduct" and "is aware that a consequence will occur in the ordinary course of events." The jurisprudence of the International Criminal Court (ICC) has adhered to the explicit limitations contained in Article 30 of the Rome Statute, thereby excluding indirect intent. The mainstream view continues to follow Article 30, limiting criminal intent to first-degree direct intent (*dolus directus* of the first degree) and second-degree direct intent (*dolus directus* of the second degree). Therefore, in the absence of clearer evidence on State practice proving that the subjective elements of war crimes universally include indirect intent, it is unfeasible, under the direct intent standard as currently applied by the ICC, to pursue the operator's criminal responsibility as a principal perpetrator based on an unlawful attack by an autonomous weapon that exceeds the operator's preset parameters.

That said, the direct intent standard is not the only possibility. Some national jurisdictions and international tribunals do recognise recklessness or *dolus eventualis* as a sufficient *mens rea* for war crimes.[27] This raises the question of whether a recklessness pathway might resolve the ICL dilemma. However, even if recklessness

---

[26] See, e.g., Mucić et al. ("Čelebići") Trial Judgment, 16 November 1998, para. 437; Blaškić Trial Judgment, para. 153; Tadić, Appeal Judgment, 15 July 1999, para. 232; Stakić Trial Judgment, 13 July 2003, para. 587; Brđanin Trial Judgement, *ibid*, para. 589; Naletelić Trial Judgement, para. 577; Kordić and Čerkez Trial Judgement, para. 341; Strugar Appeal Judgement, para. 277; Prosecutor v. Strugar ("Dubrovnik"), Appeal Judgment, IT-01-42-A, 17 July 2008, para. 270; Prosecutor v. Galić, Judgment, IT-98-29-T, 5 December 2003, para. 55.

[27] The Rome Statute is not the exclusive source for determining the mental state required for war crimes liability; customary international law may set different standards, which could arguably be accommodated through Article 30's "unless otherwise provided" clause in International Criminal Court Statute. See Gerhard Werle and Florian Jessberger, "'Unless Otherwise Provided': Article 30 of the ICC Statute and the Mental Element of Crimes under International Criminal Law", *Journal of International Criminal Justice*, Vol. 3, 2005, pp. 35-55.

is accepted as the applicable mental state, two structural problems remain. First, recklessness must attach to a specific criminal act, and without reconceiving the deployment authorization, rather than the machine's autonomous strike, as the relevant act, there is no conduct to which the commander's reckless state of mind can be connected. Second, establishing a causal link between the commander's reckless decision and the specific downstream harm requires bridging the algorithmic black box, which conventional proximate causation analysis cannot accomplish. Overcoming these structural problems requires fresh perspective and analytical framework.

Beyond this, theoretically, one might attempt to stylize LAWS as a criminal subject and seek to pursue the responsibility of the humans behind them through "Joint Criminal Enterprise" (JCE) or "Command Responsibility." [28] Even disregarding the previously mentioned consensus against anthropomorphizing LAWS into legal subjects,[29] this argumentative path faces fundamental obstacles.

Faced with algorithmic entities possessing a high degree of autonomous agency, epistemological barriers render the commander unable to penetrate the black box of LAWS to "know" exactly when and where the system will spill over.[30] JCE requires participants to share a "common purpose" and a "meeting of minds".[31] The *ad hoc* tribunals further clarified that a defendant need not have foreseen an excess act by a group member as inevitable; it suffices that the act was a "possible consequence" of executing that common purpose. In this context, the conundrum lies in how to explain that both the autonomous weapon and the operator or their commander recognise the existence of a shared "common purpose," as well as the autonomous weapon's volition to advance such a common purpose.

The subjective elements for establishing command responsibility are that the defendant, acting as a superior, knew or had reason to know that their subordinates were committing or about to commit crimes. [32] Both customary law and ICC jurisprudence confirm that the legal standard for determining a superior-subordinate relationship is whether the commander personally exercises "effective control" over

[28] For a recent comprehensive analysis of JCE and command responsibility challenges in the context of autonomous weapon systems, see Antonio Coco, "Modes of Liability for AI-Enabled Crimes in International Criminal Law", *International Law Studies*, Vol. 107, 2026, pp.228-255.
[29] Regarding this methodological approach, a possible counter-argument exists: although the law does not recognise autonomous weapons as having subject status, for the necessity of applying existing international criminal law and State responsibility rules, autonomous weapons are treated as fictive subjects. The purpose is not to pursue the responsibility of the fictive subject, but to focus on pursuing the responsibility of individuals and States as subjects of international criminal law.
[30] Prosecutor v. Tadić, Appeals Judgment, para.228.
[31] See Tadić Appeals Judgment, para. 227-228, and Prosecutor v. Brđanin, Trial Judgment, para. 364.
[32] See Article 28, Rome Statute; see also Prosecutor v. Bemba, ICC-01/05-01/08, Trial Judgment, 21 March 2016, paras. 170-171.

the subordinate.[33] The paradox here is: if the commander can effectively control the autonomous weapon as a fictive "subordinate," it implies that there is no possibility of the autonomous weapon breaching the operator's settings to execute an unlawful attack. Furthermore, in existing mechanisms, a commander's omission manifests as a failure to prevent crimes by subordinates and a failure to punish crimes already or currently being committed by subordinates. Both prevention and punishment are premised on the counterpart being human.[34] Particularly regarding punitive measures, it is hard to imagine LAWS being subjected to human criminal punishment as a "subordinate" – for example, turning off a machine does not equate to a criminal penalty.

D. The Inapplicability of Possible Dogmatic Defenses

Could highly autonomous AI weapon systems be downgraded conceptually, treating the unlawful consequences caused by their spillover as a "weapon malfunction" or stylizing them as a "mistake of fact," thereby subsuming them within the law's traditional interpretative framework? The answer to both questions is negative, for the following reasons.

First, agency spillover is by no means a traditional weapon malfunction or mechanical failure. In the law of armed conflict, traditional weapons are deterministic systems. If errors such as a bomb deviating from its trajectory or a fuze failing occur, they are regarded as accidental events caused by the laws of physics or material fatigue. As long as the weapon itself is lawful and the commander has fulfilled the duty of reasonable care prior to launch, such pure physical failures are typically treated as accidents on the battlefield and are generally not considered acts violating IHL that give rise to State responsibility or individual criminal responsibility.[35]

Equating the agency spillover of LAWS with a weapon malfunction ignores the ontological characteristics of machine learning technology. An ASE is the optimal solution autonomously derived by the system based on the probabilistic calculation of environmental data by its built-in algorithms, while its hardware is functioning normally and its sensors have no faults. For example, if LAWS, exceeding the rules of engagement set by human operators, identify a civilian

---

[33] See Prosecutor v. Bemba, ICC-01/05-01/08, Trial Judgment, paras. 183-184; Prosecutor v. Delalić et al. (Čelebići), Trial Judgment, para. 378.

[34] See generally Kai Ambos, "Superior Responsibility", in Antonio Cassese *et al.* (*eds*), *The Rome Statute of the International Criminal Court: A Commentary*, Oxford University Press, Oxford, 2002, pp. 849-850.

[35] It is worth noting that in relevant jurisprudence of international criminal law, the issue of the margin of error of weapons has been discussed — that is, within what margin an attack's error falls within the tolerance of the attacking party's pursuit of military advantage, and beyond which the attack error becomes an unlawful bombardment, thereby constituting evidence of a suspected war crime. See Prosecutor v. Gotovina et al., IT-06-90-A, Trial Judgments, para.1898; Prosecutor v. Gotovina et al., IT-06-90-A, Appeals Judgments, para.58.

communication tower of a specific structure as an enemy radar and destroy it, it is not because the system is malfunctioning. Rather, its algorithmic agent exercised its endowed functional agency when executing feature matching.

Second, agency spillover should not be simply stylized as a combatant's "mistake of fact." Article 32 of the Rome Statute established "mistake of fact" as a ground for excluding criminal responsibility. Its original intent was to accommodate the cognitive and psychological limitations of humans under the fog of war, such as misjudgements caused by obstructed vision or extreme fear. This is recognised as a lawful justification that negates the subjective criminal intent. Attempting to stylize the identification deviations of LAWS as a machine's mistake is a quintessential anthropomorphizing fallacy. The premise of a mistake is that the subject possesses consciousness and intentionality, capable of generating a misalignment between subjective cognition and objective reality. However, as an "agent without intention,"[36] LAWS do not experience fear, nor are they psychologically affected by battlefield pressure.

Concurrently, the individual commander cannot use this as a defense either. In an ASE, the commander does not directly observe the specific target of engagement. Instead, the cognitive and decision-making processes are substantively outsourced to the machine. Here what the commander faces is not a mistake of fact regarding a specific battlefield reality, but rather a blind reliance on the capability boundaries of the proxy system.

III. The Analytical Perspective of Human-Machine Hybrid Agency

The preceding dogmatic analysis demonstrates how traditional international law falls into a rupture of attribution and evaluation when confronting LAWS in ASE. One of the reasons is that its underlying jurisprudence has experienced a dislocation in both the cognition and handling of a novel type of agency.[37] Since the agency spillover of machines is the direct cause of the dilemma in legal application, the deep excavation and reconstruction of agency as a core jurisprudential concept may serve as a feasible path to break this deadlock.

A. The Evolution of Agency in Legal History

In orthodox legal and moral philosophy, agency has long been deeply bound to free will, moral perception, and intention, being regarded as the exclusive privilege of

[36] See Ian Ayres and Jack M. Balkin, "The Law of AI is the Law of Risky Agents Without Intentions", *University of Chicago Law Review Online*, 2024, available at: https://lawreview.uchicago.edu/online-archive/law-ai-law-risky-agents-without-intentions.

[37] Whether it is IHL, ARSIWA, or ICL, the construction of their legal functions is invariably built upon the single closed loop of the human moral agency.

human beings, or fictive legal persons composed of humans.[38] However, a concise review of legal evolutionary history reveals that the legal system has not exclusively possessed the capacity to handle human agency from beginning to end. Faced with historical impacts of agency originating from heterogeneous sources that exceeded human micro-control, various sub-branches of law have successfully iterated their analytical frameworks on multiple occasions.

One classic example is biological agency. Among the earlier non-human agentic entities addressed in legal history were animals. When a warhorse was startled or a vicious dog injured someone, the animal exhibited a behavioural spillover that deviated from its owner's will. The law never consequently endowed animals with legal personhood or moral subject status, nor did it simply treat them as a mechanical malfunction akin to a broken carriage axle. Instead, by objectively recognizing the animal's instinctive agency, tort law shifted the focus of attribution from seeking the owner's micro-operational fault to controlling the activation of the source of danger. This logic of addressing non-human agency through risk allocation and strict liability is a classic paradigm for the law to handle systems that escape micro-control.[39]

Artefactual agency, or group agency, serves as another example.[40] With the rise of the Industrial Revolution and modern commerce, collective entities such as corporations and organizations demonstrated an agency surpassing that of any single natural person. Through doctrines like vicarious liability, the law recognised that a principal must bear the costs for the autonomous acts of an agent within the scope of authorization. Its jurisprudential essence is: when you expand your own power and pursue your own interests by authorizing an entity with independent capacity for action, you must bear the responsibility for the structural spillover risks it generates.

The law's accommodation of agency is not absolutely premised on the entity possessing human consciousness. When a novel type of agency disrupts the existing

---

[38] See, e.g., Michael Bratman, *Intention, Plans, and Practical Reason*, Harvard University Press, Cambridge, MA, 1987. For the classical link between moral agency and free will, see Immanuel Kant, *Groundwork of the Metaphysics of Morals* (1785), 4:446-448. On the legal tradition's reliance on this presumption, see John Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge University Press, Cambridge, 1998, Ch. 1.

[39] For the Roman law origins, see Justinian, *Digest*, Book 9, Title 1 (*De pauperie*), establishing the *actio de pauperie* for harm caused by animals acting *contra naturam sui generis*. For a comparative analysis of strict liability regimes for dangerous animals across legal systems, see Cees van Dam, *European Tort Law*, 2nd ed., Oxford University Press, Oxford, 2013, pp. 371-380. On the structural parallel between animal liability and emerging AI liability frameworks, see Simon Chesterman, *We, the Robots? Regulating Artificial Intelligence and the Limits of the Law*, Cambridge University Press, Cambridge, 2021, pp. 158-162.

[40] In legal philosophy, a corporation is viewed as a typical form of Group Agency. The law resolves its spillover effects by granting it derivative rights and imposing vicarious liability. This jurisprudential logic of imposing obligations on non-human entities can be utilized to inspire the regulation of AI. See Christian List, "Group Agency and Artificial Intelligence", *Philosophy & Technology*, Vol. 34, 2021, pp.1213-1242.

order, the corresponding law successfully achieves a re-anchoring of complex causal chains by separating factual capability of action from subjective moral intent and timely updating principles of attribution, with risk liability and vicarious liability being prime examples.

B. The Human-Machine Hybrid Agency

AI has brought about a new form of agency.[41] Drawing on recent research in the philosophy of information and AI law, AI agency can be defined as a factual/functional agency that is independent of human consciousness and moral intent, capable of independently evaluating inputs based on underlying machine learning architectures in open, unstructured physical environments, and autonomously generating intervention strategies and physical consequences.[42] This definition is equally applicable to characterizing the AI-based agency of LAWS.

At this juncture, the law can no longer simply treat LAWS as passive tools or objects, nor can it analyse complex engagements solely through the narrow lens of human subject agency. We must consciously accomplish an ontological perspective shift in legal analysis. On the modern battlefield where LAWS are deployed, focusing solely on singular human agency or pure machine agency is one-sided; the complete analytical object faced by the law is a "Hybrid Agency." It is interwoven from the human (moral) agency embedding the original intent of engagement and initial activation settings, and the functional agency manifested through the machine's subsequent autonomous execution.

To demonstrate why hybrid agency is the necessary conceptual adjustment, we can see how each of the three competing framings produces a specific analytical failure. The first treats LAWS as a **pure tool**, i.e. the human is the sole agent, the weapon is an instrument. In the spillover scenario, this forces the law to evaluate the human's conduct at the moment of the strike, where the human factually had no cognitive access to what happened. We must either attribute to the human a mental state they did not possess, thus distorting *mens rea*; or conclude that no agent is responsible, thus creating an accountability gap. The second treats LAWS as an **autonomous agent** with independent responsibility. This fails not only because LAWS lack legal personhood, but because the machine's agency is constituted by upstream human decisions during design, training, and deployment parameters setting. Treating it as independent severs this constitutive relationship, making human decisions legally invisible. The third and most sophisticated alternative treats the engagement as **sequential agency**, i.e. human acts first, machine acts second, each analysed separately. This fails because it creates a fatal attribution gap at the

[41] L. Floridi, above note 4, p.30.

[42] Floridi explicitly proposes the "Multiple Realisability of Agency" thesis, advocating for defining AI as an "Agency without Intelligence" that requires no human intelligence, intent, or mental state, and arguing that the law should regulate its "functional agency" as an entity agent. See L. Floridi, above note 4, p.30.

seam. The human's agency has ended after they authorized deployment, but the machine's harmful act has not yet occurred, and the causal chain passes through a void attributable to neither. Moreover, IHL evaluates "attacks" holistically; the proportionality principle requires balancing with respect to the whole operation, not isolated segments. Hybrid agency avoids all three failures. It is not a metaphysical claim but a *legal-analytical* claim: when a human delegates cognitive and executive functions to a machine with autonomous behavioural capacity, the legally relevant agent is the composite system.

A boundary question arises: if the commander and the machine form a hybrid, why not extend the concept to include the system's developer, trainer, or procurement officer? The answer lies in operational authorship. Hybrid agency encompasses those actors whose cognitive and executive contributions are functionally indivisible at the level of the specific operational event being evaluated. The commander is included because they authorise the specific deployment in the specific operational context. The machine is included because it executes the tactical decisions within that operational envelope. Upstream contributors such as developers and trainers shaped the system's general capabilities, but they did not author this operation. Their potential liability runs through separate legal channels possibly like product liability or contractual liability, not through the hybrid agent concept.

Introducing the perspective of hybrid agency to analyse the legal issues of LAWS bears significant jurisprudential meaning. It inspires international law scholars examining the legal issues of ASE to stop fruitlessly searching for specific human malice within the machine's black box, or forcefully using "mistake of fact" to cover up the machine's computational autonomy. The human commander provides the normative framework, initial target settings, and authorization for the engagement, while the AI agent fills in the specific spatiotemporal trajectory and target selection. Internally, the agency of these two stages is relatively independent, yet there exist complex, difficult-to-explain or even incomprehensible linkages. However, for external analysis, they exhibit a hybrid agency that can be addressed holistically. Therefore, the determination of whether IHL obligations have been violated can no longer divisibly evaluate the human's button-pressing action in the rear or the machine's physical destruction at the front; instead, the "human-machine hybrid" must be systematically evaluated as an indivisible cognitive and operational unit.

C. The Principal-Agent Analytical Lens and Its Derivative Tools

The essence of hybrid agency is that humans, in order to acquire computational or reaction efficiency beyond their own limits, voluntarily delegate certain cognitive and executive functions to machines. This structural delegation of power and division of function exhibits features familiar from the Principal-Agent (P-A)

relationship in economics and organizational behaviour,[43] making it a valuable analytical lens for illuminating the internal structure of the human-machine hybrid agency.[44] The P-A lens possesses a unique efficacy in addressing the agency conundrum: it can capture the information asymmetry and alignment problem that inevitably accompany agency relationships. In the algorithmic context, the human principal, constrained by cognitive limits, cannot penetrate the algorithmic black box in real time; meanwhile, the machine, acting as the agent, lacks moral intuition and merely pursues the mechanical maximization of its reward function. When environmental data spills over the preset distribution, a phenomenon termed "out-of-distribution" (OOD) in machine learning,[45] the factual agency exhibited by the machine is highly prone to deviate from the human's initial intent. This deviation is not a mere mechanical tool malfunction, but an agency cost inevitably endogenous to the principal-agent structure. Thus, relying on the principal-agent lens allows for the introduction of a mature set of conceptual tools and theoretical logic regarding power delegation, information asymmetry, and the assumption of agency risks and liabilities for subsequent analysis.

However, as an analytical lens, the principal-agent perspective must be combined with domain-specific legal reasoning before it can resolve concrete legal issues. This is because when different legal normative systems confront the same agentic entity, the core questions they pursue and their attribution logics are starkly different. Just as in traditional domestic law, when facing the agency structure of a corporate legal person, contract law tends to apply apparent agency to protect the appearance of transactions, whereas tort law tends to apply vicarious liability to allocate spillover risks. Similarly, facing a human-machine hybrid agency network, we must derive specific analytical tools based on different legal scrutiny objectives.

The first is the Joint Cognitive System model developed based on cognitive science,[46] applicable to evaluating the legality of the behavioural process and decision-making chain. In a highly automated agency network, the high-frequency data processing of machines constitutes a filter for human perception, and humans often lose substantive interventional capacity at the tactical edge. This model posits

---

[43] The Principal-Agent theory in economics and common law has proven to be an effective tool for analyzing the AI black box and the alignment problem. Because AI lacks intrinsic "single-minded loyalty", hidden information and hidden action lead to endogenous agency costs, which naturally aligns in law with the applicative logic of *Respondeat Superior* (vicarious liability). See Anat Lior, "AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy", *Mitchell Hamline Law Review*, Vol. 46, Iss. 5, 2020, pp.1043-1071. Also see Noam Kolt, *Governing AI Agents*, arXiv preprint arXiv:2501.07913, 2025, pp.18-21.

[44] The P-A lens is employed here not as a full theoretical import but as a structural analogy: in the classical framework, the agent has private interests that diverge from the principal's, producing agency costs. LAWS lack private interests in this sense; the divergence between the machine's behaviour and the human's intent is epistemic rather than motivational.

[45] See Andrew J. Lohn, *Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing out-of-distribution Performance*, arXiv preprint arXiv:2009.00802, 2020, pp.5-6.

[46] See Luciano Floridi, "Faultless Responsibility: On the Nature and Allocation of Moral Responsibility for Distributed Moral Actions", *Philosophical Transactions of the Royal Society*, Vol. 374, 2016.

that, in such scenarios, AI is no longer a passive executor independent of the human, but an extension of human cognitive organs; the true bearer of obligations should be defined as the human-machine joint cognitive system. The rationality of this model lies in its establishment of distributed causality. Since the edge decision-making is already indivisible, the law's scrutiny of compliance and feasibility must be shifted leftwards on the timeline, distributed across the entire lifecycle stages of the system, including design, training, and the establishment of macro-constraints.

The second is the Risk Allocation model, which absorbs the jurisprudence of traditional strict liability, applicable to the objective attribution of harmful consequences and loss compensation.[47] Within the agency structure, the factual agency of the agent often produces spillover consequences exceeding the principal's specific instructions. To prevent the principal from exploiting this agency spillover to sever the causal link and evade responsibility, this model analogizes systems with algorithmic autonomy to dangerous animals possessing biological instincts. The logical self-consistency of this tool is built upon the principles of non-reciprocal risk and the alignment of benefit and risk, that is benefit-risk equivalence. Since the principal, for their own benefit, releases an agentic entity with potential destructive power, the incidental damage caused by this entity based on its algorithmic instincts constitutes an inherent internal risk of the system. This model provides theoretical support for establishing absolute vicarious liability without fault and stripping the principal of any room for plausible deniability.

The third is the Reckless Delegation model,[48] focusing on the authorization decision itself, applicable to assessing the individual subjective culpability and punitive evaluation. Noting the legal principle of *nullum crimen sine lege* (no crime without law), this model does not simply advocate mimicking the principal-agent liability models of civil and commercial law to impose the consequences of suspected war crimes from LAWS' agency spillover engagements upon human subjects. Rather, it argues for shifting the anchor of evaluation from the transgressive acts at the agent's behavioural end to the principal's frontend authorization and decision-making behaviour. This tool demonstrates that if a principal, fully aware that the introduced black-box agent poses a systemic risk of misalignment in complex environments, still delegates lethal force to it without physical limitations or architectural constraints, this decision itself embodies an extreme reckless disregard for the lawful rights of others. This interpretative logic can convert the direct intent

---

[47] The risk allocation approach to AI liability draws on the tradition of strict liability for ultrahazardous activities. See Anat Lior, "AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy", *Mitchell Hamline Law Review*, Vol. 46, Iss. 5, 2020, pp.1043-1071; see also Gerhard Wagner, "Robot Liability", in Sebastian Lohsse *et al. (eds)*, *Liability for Artificial Intelligence and the Internet of Things*, Nomos, 2019, pp. 27-62.

[48] On the concept of reckless risk-creation as a basis for criminal liability in the context of autonomous systems, see I. Ayres and J. Balkin, above note 36. For the broader doctrinal foundations of shifting *actus reus* to the deployment decision, see Noam Kolt, Governing AI Agents, arXiv preprint arXiv:2501.07913, 2025, pp. 18-21.

for specific harmful outcomes into recklessness in creating an impermissible agency risk, thereby providing a viable legal reasoning path for punishing irresponsible algorithmic applications while firmly upholding the baseline of subjective attribution.

To avoid collapsing into *de facto* strict criminal liability, this model incorporates a culpability threshold. The delegation is reckless when the commander knew or willfully disregarded a substantial and unjustifiable risk of agency spillover in the specific operational context *and* failed to implement available architectural constraints that would have materially reduced that risk. If a commander implemented every reasonably available safeguard, such as the adversarial testing, geofencing, target-class restrictions, override mechanisms, and still experienced spillover, he or she has not engaged in reckless delegation. That is treated as a reasonable decision with a bad outcome, not a culpable one.

IV. Transcending the Legal Dilemmas in Agency Spillover Engagements

The perspective of human-machine hybrid agency and the principal-agent analytical lens provide the possibility to reactivate existing law. The fundamental utility of this theoretical perspective is that it elevates the analytical unit of legal evaluation from isolated tactical actions to a systemic network of authorization and decision-making. It reveals that the agency spillover of machines is not an incomprehensible physical malfunction, but an endogenous, structural agency cost inevitably arising after the delegation of power. Based on this reconceptualisation, we need not search in vain for human micro-faults within the machine's algorithmic black box, nor do we need to endow the machine with a fictive legal personhood. Instead, we can utilize specific analytical tools to attempt to reshape the connotations of old legal concepts through evolutionary and teleological interpretation, without departing from the existing positive law (*lex lata*) framework.[49] In doing so, we could achieve a substantive transcendence of the three major legal dilemmas within the realm of legal hermeneutics.

A. Dogmatic Updating of the Application of Substantive IHL Rules

The primary IHL dilemma revealed previously is that the high-frequency computational agency of machines deprives humans of the physical possibility of intervention at the tactical end, resulting in the failure of both the feasibility requirement of the precautionary principle and the subjective balancing required by the principles of distinction and proportionality. The Joint Cognitive System analytical tool may prove helpful to transcend this dilemma.

[49] A methodological note is warranted here. This paper's commitment to *lex lata* is not merely strategic but substantive. The primary obligation of legal scholarship confronting a new technological reality is to test rigorously how far existing law can stretch through interpretation before concluding that new law is needed.

From the perspective of legal hermeneutics, API Article 57 requires the attacker to take "all feasible precautions" to verify targets and control incidental damage. In the context of traditional weapons, the literal meaning of "feasible" usually refers to visual confirmation or radar calibration at the very moment of engagement. However, in a human-machine hybrid agency network, because the terminal decision has been substantively delegated to the agent, demanding humans to fulfill verification and suspension obligations in millisecond-level engagements is physically no longer "feasible." The ultimate purpose of IHL is to protect civilians and limit unnecessary suffering. To prevent this normative purpose from being circumvented by technology, the conceptual extensions of feasibility and attack must be dynamically expanded. Based on the distributed causality established by the Joint Cognitive System, since human cognition and machine algorithms are indivisible during the execution phase, the law's scrutiny of the obligations of distinction and proportionality must undergo a systemic "shift left" – i.e. to expand the compliance span to *ex ante* lifecycle – on the timeline.

This means that feasible precautions under interpretative legal theory are no longer confined to the moment the trigger is pulled but are lawfully expanded to the entire lifecycle of the system.[50] For example: Did the commander and operator conduct sufficient adversarial testing (red-teaming) on out-of-distribution data for the specific theater of operations prior to deployment? Did they establish strict geofencing and forced-sleep time windows for the system in accordance with the ROE? Only if the commander has fulfilled the compliance obligations of the joint system at the frontend stages of system constraint setting and deployment authorization can they be deemed to have fulfilled the substantive obligations of IHL. Conversely, neglecting to establish architectural constraints at the frontend, thereby leading to agency spillover, would constitute a suspected violation of the principles of distinction or proportionality.

B. Evolutionary Interpretation of the Attribution Rules in State Responsibility

In the *Namibia* and *Costa Rica v. Nicaragua* cases,[51] the International Court of Justice (ICJ) established that generic terms with evolutionary meanings should be interpreted in accordance with the legal and factual realities at the time of

[50] Recent international practice emphasizes that human agency and IHL compliance cannot be evaluated solely at the moment of engagement, but must permeate the entire lifecycle of AI (from research and development, testing, and procurement to deployment and post-use assessment). See The AutoPractices Team (Ingvild Bode *et al.*), *Strengthening Human Agency in the Military Domain* (Center for War Studies, 2026), pp. 21-27. See also the CCW GGE Rolling Text on LAWS, 18 December 2025, CCW/GGE/LAWS, especially Box IV on human responsibility across the lifecycle; and also Jonathan Kwik, "Iterative Assessment for Military Artificial Intelligence (AI) Systems", in Bérénice Boutin, Taylor Kate Woodcock, Sadjad Soltanzadeh (*eds*) *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence*. T.M.C. Asser Press, The Hague, 2026.

[51] ICJ, *Legal Consequences for States of the Continued Presence of South Africa in Namibia (South West Africa)*, Advisory Opinion, ICJ Reports 1971, p. 31, para. 53; ICJ, *Dispute regarding Navigational and Related Rights (Costa Rica v. Nicaragua)*, Judgment, ICJ Reports 2009, p. 213, para. 64.

application. By introducing the analytical perspective of hybrid agency, when conducting legal attribution arguments under State responsibility law for suspected unlawful acts arising from LAWS in ASE, the following interpretative updates can be made; and it is useful to present them in order of doctrinal conservatism.

The most conservative textual anchor is Article 7 of ARSIWA, which provides that the conduct of an organ of a State shall be considered an act of the State "even if it exceeds its authority or contravenes instructions." The human operator or commander who deploys LAWS is a State organ under Article 4. Under hybrid agency, the human-machine system is the operational unit through which this State organ exercises military force. When the system produces an ASE, the resulting conduct, namely the unauthorized strike, is conduct that *exceeds the authority* given to the system by the human component. Article 7 captures precisely this situation. Without hybrid agency, the State could argue that the organ's own conduct, i.e. pressing the deploy button, was lawful, and the unlawful conduct was the machine's alone. With hybrid agency, the engagement is indivisible: the organ acted through the machine, and the machine's spillover remains part of the organ's *ultra vires* conduct.

Alternatively, by linking the suspected unlawful conduct holistically to the hybrid agency without delving into its internal agency allocation mechanisms, the human-machine hybrid as a whole can be treated as the "person or entity" exercising elements of governmental authority under Article 5 of ARSIWA. There is currently no general framework in international law regulating State responsibility for inanimate objects, with only specific regimes existing, such as those applicable to space objects or transboundary harm caused by hazardous activities.[52] However, as previously noted, an authorized entity can encompass a wide range of forms, provided it is empowered by the internal law of the State. In other words, within the operational margin of treaty interpretation, there is an opportunity to achieve attribution justification through this path.

Lastly, the interpretation of ARSIWA Article 8 can also be navigated. Within the interpretative framework of existing law, the "effective control" standard established by the ICJ in classic jurisprudence typically requires the State to exercise micro-command over specific tactical operations. Confronted with LAWS, we could apply a teleological interpretation to theoretically upgrade the concept of "control" in Article 8. Since the State, acting as the principal through its commanders and operators, voluntarily delegates lethal force to a machine agent to gain an asymmetric military advantage, this macro "systemic authorization" inherently constitutes a form of "control" consistent with the normative intent of ARSIWA. The spillover attacks emerging from the machine based on its reward function as agency costs belong to the endogenous risks within the inherent probabilities of this hazardous

[52] See Magdalena Pacholska, "Military Artificial Intelligence and the Principle of Distinction: A State Responsibility Perspective", *Israel Law Review*, 2023, pp.20-21.

system. Therefore, interpreting such acts as "conduct under the direction or control of a State" aligns with the normative intent of ARSIWA, thereby blocking the State's attempt to use the technological black box to sever the causal chain.

C. Reinterpreting the Subjective Elements of Individual Accountability in ICL

In the realm of ICL, the pursuit of individual criminal responsibility is obstructed by the stringent subjective requirements of Article 30 of the Rome Statute, namely, the commander's lack of specific intent and knowledge regarding the particular civilian casualties caused by the machine's spillover. Utilizing the Reckless Delegation model, we may complete the closed loop of criminal dogmatics.[53] Simply put, in positive law, we do not attempt to prove that the commander had direct intent toward the specific civilian ultimately bombed by mistake, which contradicts the facts or presents legal dilemmas; instead, we propose treating the commander's deployment act as the *actus reus* under scrutiny.

Utilizing the Reckless Delegation model, the anchor of evaluation shifts from the unlawful moment of engagement to the authorization decision. When a commander knows full well that a certain type of LAWS, acting as a black-box agent, poses an extremely high risk of misalignment and spillover in a complex electromagnetic urban theater, yet still signs the deployment order without implementing effective physical geofences in pursuit of tactical advantage, then, this decision arguably satisfies Rome Statute Article 30's "knowledge" requirement. Article 30(2)(b) defines knowledge as awareness that "a consequence will occur in the ordinary course of events." A commander who deploys a system with a known risk profile of autonomous misidentification in the given operational context is aware that civilian harm is a statistically predictable consequence of that deployment. In other words, it may not occur in every instance, but rather "in the ordinary course of events" across deployments of this type in similar conditions.

Should the above interpretation be considered too expansive, an alternative basis exists. Article 30 opens with the caveat "unless otherwise provided," acknowledging that specific crimes may carry different *mens rea* requirements under other sources of law. Customary international law, as reflected in the practice of international criminal tribunals including the ICTY, recognises that *dolus eventualis* or recklessness may satisfy the *mens rea* for certain war crimes.[54] Where this is the case, the recklessness standard, which the deployment decision described above clearly meets, applies independently of Article 30's default threshold.

---

[53] As identified in Section II.C, even where recklessness is accepted as the applicable mental state, two structural problems must be overcome: the *actus reus* must be reconceived, and the causal link must be established through the algorithmic black box. The Reckless Delegation model addresses both.

[54] Jurisprudentially, regarding "risky agents without intentions," the law may abandon the search for the underlying micro-level subjective intent and instead determine the recklessness or gross negligence of the deployer through objective standards of risk control. See I. Ayres and J. Balkin, above note 36.

A further challenge is that criminal law requires specificity in the risk the defendant foresees. In other words, a vague sense that "something might go wrong" is insufficient to establish *dolus eventualis*.[55] The hybrid agency framework helps address this in two ways. First, the Joint Cognitive System model specifies *what* the commander should know. The system's documented failure modes, its performance profile in adversarial testing, and the gap between training distribution and operational environment may all be included. This is not diffuse unease but technically specific risk awareness. Second, the Reckless Delegation model defines *what kind of risk* counts: agency spillover producing civilian harm from autonomous misidentification in the specific type of operational environment where the system is deployed. The commander's awareness is assessed against the system's known risk profile, which is a concrete, documentable set of facts, rather than against a general anxiety about AI unpredictability. That said, borderline cases will inevitably arise, and the precise threshold of risk specificity will require judicial development.

V. Conclusion

The rapid evolution of military artificial intelligence has pushed IHL to an unavoidable ontological crossroads. The ASE declares the localized failure of traditional legal application built upon the presumption of a pure human moral subject. The key to breaking this structural crisis is neither mechanically reducing the algorithm to a mere tool malfunction nor taking a radical leap to endow it with personhood, but rather facing head-on the human-machine hybrid agency in modern armed conflict.

By employing the Principal-Agent relationship as an analytical lens and developing three derived analytical tools, this paper demonstrates that *Lex Lata* still possesses robust resilience to regulate algorithmic violence. Through the Joint Cognitive System, IHL compliance obligations are anchored to the system's lifecycle compliance. Through the Risk Allocation model, State attribution in ARSIWA is extended to the assumption of systemic agency costs. Through the Reckless Delegation model, the punitive focus of ICL can be redirected to the decision-making hub that abuses black-box algorithms.

This series of jurisprudential reconstructions directly responds to and empowers the ongoing governance processes of global military AI. It provides a theoretical basis for implementing meaningful human control in the Geneva CCW GGE negotiations,[56] implying that the failure of micro-control must be replaced by

---

[55] See Jonathan Kwik, "The Conceptual Roots of The Criminal Responsibility Gap in Autonomous Weapon Systems", *Melbourne Journal of International Law*, Vol. 24, 2023, pp. 20-21.

[56] See Chair's summary – Second 2025 session of the GGE on LAWS - Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, 20 October 2025, available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional_ Weapons_-

architectural red lines. Concurrently, it injects a hard-law accountability logic into the lifecycle governance advocated by the GC REAIM initiative.[57]

Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_ Systems_(2025)/CCW-GGE.1-2025-WP.9_-_Chair's_summary.pdf

[57] See GC REAIM, "*Responsible by Design: Strategic Guidance Report on the Risks, Opportunities, and Governance of Artificial Intelligence in the Military Domain*", September 2025, available at: https://hcss.nl/wp-content/uploads/2025/09/GC-REAIM-Strategic-Guidance-Report-Final-WEB.pdf